CLAIMS

1. Apparatus for providing high-performance, scaleable data processing and storage services to a client from a plurality of resources, comprising

an access interface module which receives requests for service from the client and selects a subset of the plurality of resources to provide the requested service and distribute the workload across the plurality of resources; and

a switch fabric for temporarily connecting the access interface module to the selected subset of the plurality of resources for providing the service to the client.

2. The apparatus of claim 1 wherein the access interface module selects the subset of the plurality of resources based on the relative demand placed on the subset of resources.

3. The apparatus of claim 1 wherein the switch fabric comprises a control switch fabric for transferring control information and a separate data switch fabric for transferring data.

4. The apparatus of claim 3 wherein the control switch fabric is optimized for transferring control information and the data switch fabric is optimized for transferring data.

5. The apparatus of claim 3 wherein the request for service includes control information and data and wherein the access interface module separates the control information and the data and transfers the data to the selected subset of resources over the data switch fabric.

6. The apparatus of claim 3 wherein the data switch fabric comprises a non-blocking crossbar switch for data transfer and the control switch fabric comprises an Ethernet switch for control information transfer.

7. The apparatus of claim 1 further comprising a resource module connected to the plurality of resources for generating preallocation information that preallocates services from the plurality of resources in order to evenly distribute a workload across the plurality of resources.

8. The apparatus of claim 7 wherein the switch fabric connects the access interface module to the resource module so that the resource module can transfer the preallocation information to the access interface module.

9. The apparatus of claim 8 wherein the access interface module selects a subset of the plurality of resources based on the preallocation information.

10. The apparatus of claim 1 wherein the access interface module comprises a data memory which temporarily stores information transferred between the access interface module and the selected subset of the plurality of resources.

11. The apparatus of claim 1 further comprising a plurality of access interface modules each access interface module receiving service requests from a plurality of clients.

12. A disk-based storage system for providing high-performance, scaleable storage services to a client from a plurality of disks, comprising

a disk interface module connected to the plurality of disks for controlling data stored on the plurality of disks;

a host interface module which receives requests for storage service from the client and selects a subset of the plurality of disks to provide the requested storage and distribute the workload across the plurality of disks; and

a switch fabric for temporarily connecting the host interface module to the resource module for providing the storage service to the client.

13. The storage system of claim 12 wherein the switch fabric comprises a control switch fabric optimized for transferring control information and a separate data switch fabric optimized for transferring data.

14. The storage system of claim 13 wherein the request for service includes control information and data and wherein the host interface module separates the control information and the data and transfers the data to the selected subset of resources over the data switch fabric.

15. The storage system of claim 13 wherein the data switch fabric comprises a non-blocking crossbar switch for data transfer and the control switch fabric comprises an Ethernet switch for control information transfer.

16. The storage system of claim 12 wherein the disk interface module generates preallocation information that preallocates physical storage in the plurality of disks in order to evenly distribute data across the plurality of disks.

17. The storage system of claim 16 wherein the physical storage in the plurality of disks is divided into zones and the disk interface module preallocates selected zones to frequently-accessed data, wherein the selected zones are selected in order to decrease disk access time.

18. The storage system of claim 16 wherein the host interface module logically maps data items to be stored into allocation units preallocated to the host interface module by the disk interface modules.

19. The storage system of claim 12 wherein the host interface module comprises a first data memory and the resource module comprises a second data memory and wherein the first and second data memories temporarily store information transferred between the host interface module and the disk interface module.

20. The storage system of claim 12 further comprising a plurality of host interface modules, each host interface module receiving service requests from a plurality of clients.

21. The storage system of claim 12 further comprising a plurality of disk interface modules connected to the plurality of disks.

22. The storage system of claim 21 wherein the host interface module directs writes to disk interface modules of data based on the relative demand placed on those modules without regard to either the contents of associated disks or the location of other copies of the data.

23. The storage system of claim 22 further comprising a metadata module that is connected to the switch fabric and assigns each data object an identifying handle.

24. The storage system of claim 23 wherein the host interface module and the metadata module store metadata information associated with each data object in the plurality of disks separately from, and independently of, data associated with each data object.

25. The storage system of claim 24 wherein the handle contains the location of the metadata in the plurality of disks.

26. The storage system of claim 23 wherein the metadata module is dedicated exclusively to fetching, caching and manipulating file and directory attributes of each data object thereby allowing data paths to be optimized to maximize data flow.

27. The storage system of claim 12 wherein the disk interface module preallocates storage in the plurality of disks in allocation units.

28. The storage system of claim 27 wherein the host interface module selects a subset of the plurality of disks from the preallocated storage.

29. The storage system of claim 27 wherein the host interface module reads a file from the plurality of disks and writes the file to disk in a single allocation unit.

30. The storage system of claim 27 wherein the disk interface module comprises a data memory for temporarily storing a data file that is to be written to disk and a parity generator for generating parity information on the data file stored in the data memory prior to the transfer of the data file to the disk.

31. The storage system of claim 30 wherein data and its associated parity are stored on disk in their entirety in order to avoid reading previously stored data to determine data modifications and re-calculating parity.

32. The storage system of claim 12 wherein the host interface module stores data to disk without regard to where any earlier versions of that data were previously stored and without regard to the file and file system to which the data belongs.

33. The storage system of claim 12 wherein the data to be stored on the plurality of disks is arranged in a plurality of data pages and wherein the storage system comprises a reference counter for maintaining reference counts on each data page so that unused disk space can be readily identified.

34. The storage system of claim 12 in which the modules are interconnected using a high-speed, non-blocking, crossbar switch for transferring data.

35. The storage system of claim 12 in which separate, serial Interprocessor Communication (IPC) channels are used to transfer metadata between pairs of modules, thereby enabling the crossbar switch to be used at maximum efficiency for transferring data and allowing memory elements in the external interface to be partitioned into dedicated incoming and outgoing data buffers.

36. A fault-tolerant computer system for providing scaleable data processing and storage services to a client from a plurality of storage resources, comprising

a plurality of identical resource interface modules connected to the storage resources;

a plurality of identical access interface modules which receive requests for service from the client and select a subset of the plurality of resource interface modules to provide the requested service and distribute the workload across the plurality of storage resources; and

a switch for temporarily connecting one of the plurality of access interface modules to the selected subset of the plurality of resource interface modules for providing the service to the client, the switch being constructed in two identical halves which are interconnected so that a failure in one switch half does not make the computer system inoperative.

37. The computer system of claim 36 wherein one switch half is designated as active and the other switch half is designated as standby and the active switch half is

used to temporarily connect one of the plurality of access interface modules to the selected subset of the plurality of resource interface modules.

38. The computer system of claim 37 wherein the active switch half and the standby switch half exchange roles if the active switch half fails.

39. The computer system of claim 36 wherein each storage resource is connected to at least two resource interface modules so that a failure in any resource interface module does not prevent access to the each storage resource.

40. The computer system of claim 36 wherein each access interface module can assume the workload of any of the plurality of access interface modules so that a failure in any access interface module can be bypassed by assigning the workload of the failed module to another of the access interface modules.

41. The computer system of claim 36 wherein each resource interface module can assume the workload of any of the plurality of resource interface modules so that a failure in any resource interface module can be bypassed by assigning the workload of the failed module to another of the resource interface modules.

42. The computer system of claim 36 wherein data is stored on one of the plurality of storage resources and the computer system further comprises a parity generator which computes a data tag, including parity information, from the data and stores the data tag on the plurality of storage resource apart from the data.

43. The computer system of claim 36 wherein data is stored on one of the plurality of storage resources and an acknowledgement is returned to the client after the data has been stored and wherein the data is stored on at least two separate storage resources before the acknowledgement is returned to the client.

44. A method for providing high-performance, scaleable data processing and storage services to a client from a plurality of resources, the method comprising

(a) providing an access interface module which receives requests for service from the client;

(b) using the access interface module to select a subset of the plurality of resources to provide the requested service and distribute the workload across the plurality of resources; and

(c) using a switch fabric to temporarily connect the access interface module to the selected subset of the plurality of resources for providing the service to the client.

45. The method of claim 44 wherein step (b) comprises selecting the subset of the plurality of resources based on the relative demand placed on the subset of resources.

46. The method of claim 44 wherein step (c) comprises:

(c1) using a control switch fabric for transferring control information; and

(c2) using a separate data switch fabric for transferring data.

47. The method of claim 46 wherein step (c1) comprises optimizing the control switch fabric for transferring control information and step (c2) comprises optimizing the data switch fabric for transferring data.

48. The method of claim 46 wherein the request for service includes control information and data and wherein step (b) comprises separating the control information and the data and step (c) comprises transferring the data to the selected subset of resources over the data switch fabric.

49. The method of claim 46 wherein step (c1) comprises using a non-blocking crossbar switch for data transfer and step (c2) comprises using an Ethernet switch for control information transfer.

50. The method of claim 44 further comprising:

(d) providing a resource module connected to the plurality of resources; and

(e) using the resource module to generate preallocation information that preallocates services from the plurality of resources in order to evenly distribute a workload across the plurality of resources.

51. The method of claim 50 wherein step (c) comprises connecting the access interface module to the resource module so that the resource module can transfer the preallocation information to the access interface module.

52. The method of claim 51 wherein step (b) comprises selecting a subset of the plurality of resources based on the preallocation information.

53. The method of claim 44 wherein step (b) comprises temporarily storing information transferred between the access interface module and the selected subset of the plurality of resources.

54. The method of claim 44 wherein step (a) further comprises providing a plurality of access interface modules each access interface module receiving service requests from a plurality of clients.

55. A method for providing high-performance, scaleable storage services to a client from a plurality of disks, comprising

(a) providing a disk interface module connected to the plurality of disks for controlling data stored on the plurality of disks;

(b) providing a host interface module which receives requests for storage service from the client and selects a subset of the plurality of disks to provide the requested storage and distribute the workload across the plurality of disks; and

(c) using a switch fabric to temporarily connect the host interface module to the resource module for providing the storage service to the client.

56. The method of claim 55 wherein step (c) comprises:

(c1) using a control switch fabric optimized for transferring control information; and

(c2) using a separate data switch fabric optimized for transferring data.

57. The method of claim 56 wherein the request for service includes control information and data and wherein step (b) comprises separating the control information and the data and step (c) comprises transferring the data to the selected subset of resources over the data switch fabric.

58. The method of claim 56 wherein step (c1) comprises using a non-blocking crossbar switch for data transfer and step (c2) comprises using an Ethernet switch for control information transfer.

59. The method of claim 55 wherein step (a) comprises generating preallocation information that preallocates physical storage in the plurality of disks in order to evenly distribute data across the plurality of disks.

60. The method of claim 59 wherein the physical storage in the plurality of disks is divided into zones and step (a) further comprises preallocating selected zones to frequently-accessed data, wherein the selected zones are selected in order to decrease disk access time.

61. The method of claim 59 wherein step (b) comprises logically mapping data items to be stored into allocation units preallocated to the host interface module by the disk interface modules.

62. The method of claim 59 wherein the host interface module comprises a first data memory and the disk interface module comprises a second data memory and wherein step (a) comprises using the first data memory to temporarily store information transferred from the host interface module to the disk interface module and step (b) comprises using the second data memory to temporarily store information received by the disk interface module from the host interface module.

63. The method of claim 55 wherein step (b) further comprises providing a plurality of host interface modules, each host interface module receiving service requests from a plurality of clients.

64. The method of claim 55 wherein step (a) comprises providing a plurality of disk interface modules connected to the plurality of disks.

65. The method of claim 64 wherein the host interface module directs writes to disk interface modules of data based on the relative demand placed on those modules without regard to either the contents of associated disks or the location of other copies of the data.

66. The method of claim 65 further comprising

(d) providing a metadata module that is connected to the switch fabric and assigns each data object an identifying handle.

67. The method of claim 66 further comprising:

(f) using the host interface module and the metadata module to store metadata information associated with each data object in the plurality of disks separately from, and independently of, data associated with each data object.

68. The method of claim 66 wherein the handle contains the location of the metadata in the plurality of disks.

69. The method of claim 66 wherein the metadata module is dedicated exclusively to fetching, caching and manipulating file and directory attributes of each data object thereby allowing data paths to be optimized to maximize data flow.

70. The method of claim 55 wherein step (a) comprises using the disk interface module to preallocate storage in the plurality of disks in allocation units.

71. The method of claim 70 wherein step (a) comprises using the host interface module to select a subset of the plurality of disks from the preallocated storage.

72. The method of claim 70 wherein step (a) comprises using the host interface module to read a file from the plurality of disks and write the file to disk in a single allocation unit.

73. The method of claim 70 wherein step (a) comprises using a data memory for temporarily storing a data file that is to be written to disk and using a parity generator to generate parity information on the data file stored in the data memory prior to the transfer of the data file to the disk.

74. The method of claim 73 wherein data and its associated parity are stored on disk in their entirety in order to avoid reading previously stored data to determine data modifications and re-calculating parity.

75. The method of claim 55 wherein step (b) comprises using the host interface module to store data to disk without regard to where any earlier versions of that data were previously stored and without regard to the file and file system to which the data belongs.

76. The method of claim 55 wherein the data to be stored on the plurality of disks is arranged in a plurality of data pages and wherein the method further comprises:

(g) maintaining reference counts on each data page so that unused disk space can be readily identified.

77. The method of claim 55 wherein step (c) comprises interconnecting the modules using a high-speed, non-blocking, crossbar switch for transferring data.

78. The method of claim 55 wherein step (c) further comprises using separate, serial Interprocessor Communication (IPC) channels to transfer metadata between pairs of modules, thereby enabling the crossbar switch to be used at maximum efficiency for transferring data and allowing memory elements in the external interface to be partitioned into dedicated incoming and outgoing data buffers.

79. A method for providing fault-tolerant scaleable data processing and storage services to a client from a plurality of storage resources, comprising

(a) providing a plurality of identical resource interface modules connected to the storage resources;

(b) providing a plurality of identical access interface modules which receive requests for service from the client and select a subset of the plurality of resource interface modules to provide the requested service and distribute the workload across the plurality of storage resources; and

(c) using a switch for temporarily connecting one of the plurality of access interface modules to the selected subset of the plurality of resource

interface modules for providing the service to the client, the switch being constructed in two identical halves which are interconnected so that a failure in one switch half does not make the computer system inoperative.

80. The method of claim 79 wherein step (c) comprises designating one switch half as active and designating the other switch half as standby and using the active switch half to temporarily connect one of the plurality of access interface modules to the selected subset of the plurality of resource interface modules.

81. The method of claim 79 further comprising:

(d) exchanging the roles of the active switch half and the standby switch half if the active switch half fails.

82. The method of claim 79 wherein step (a) comprises connecting each storage resource to at least two resource interface modules so that a failure in any resource interface module does not prevent access to the each storage resource.

83. The method of claim 79 wherein step (b) comprises providing access interface modules that can assume the workload of any other of the plurality of access interface modules so that a failure in any access interface module can be bypassed by assigning the workload of the failed module to another of the access interface modules.

84. The method of claim 79 wherein step (a) comprises providing resource interface modules that can assume the workload of any other of the plurality of resource interface modules so that a failure in any resource interface module can be bypassed by assigning the workload of the failed module to another of the resource interface modules.

85. The method of claim 79 wherein data is stored on one of the plurality of storage resources and the method further comprises:

(e) computing a data tag, including parity information, from the data; and

(f) storing the data tag on the plurality of storage resource apart from the data.

86. The method of claim 79 wherein data is stored on one of the plurality of storage resources and an acknowledgement is returned to the client after the data has been stored and wherein the method further comprises:

(g) storing data on at least two separate storage resources before the acknowledgement is returned to the client.